Revisiting Evaluation of Knowledge Base Completion Models

Pouya Pezeshkpour, Yifan Tian, Sameer Singh
Revisiting Evaluation of KB Completion

Evaluating KG Completion

Shortcomings
1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC
Revisiting Evaluation of KB Completion

Overview of Knowledge Graph Completion

Shortcomings
1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Evaluating KG Completion

Introducing YAGO3-TC
Knowledge Graphs

Shortcomings of Knowledge Graphs

Because of the way they are created:

- There are many missing facts
  
  KG Completion

- The factuality of non-existent links is unknown (open-world assumption)
  
  Use of ranking for evaluation

But, real-world application mostly care about an information being True or False and not the ranking
Knowledge Graph Completion Evaluation

- Two evaluation approaches for scoring target triple \(<S, R, O>\):

1. **Ranking Metrics:**

   ![Diagram](image)

2. **Triple Classification:**

   ![Diagram](image)

   Learning thresholds \(\tau_R\) by randomly choosing negative samples for validation data

Doesn't correspond to real-world use case in many instances
Revisiting Evaluation of KB Completion

Overview of Knowledge Graph Completion

Shortcomings
1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC
Negative Sampling

• We consider 3 different negative sampling for target triple <S,R,O>:

1. Random Sampling:

Random entity as O’ or S’
Negative Sampling

• We consider 3 different negative sampling for target triple \(<S,R,O>\):

  2. Constraint Sampling:

  Random entity as \(O'\) or \(S'\)

  Appear with \(R\)
Negative Sampling

• We consider 3 different negative sampling for target triple <S,R,O>:

3. Careful Sampling:

Choose highest score entity as O’ or S’

Never appear with R
Calibration

- Calibration study is not well defined

1. Random sampling
2. Constrained sampling
3. Careful sampling

Extremely different conclusions for different negative samplings
Semi-Inverse Relations

WN18RR

R

S ~ 90%

O

R

Ranking metrics do not reflect reasoning power

YAGO3-10

playsFor

S ~ 80%

O

isAffiliatedTo

Train data

~ 40%

1/3 of test

~0.9 MRR

~ 55 %

1/2 of test

~0.9 MRR
Careful negative sampling results in a dramatic drop

Results are around 90%
Revisiting Evaluation of KB Completion

Overview of Knowledge Graph Completion

Shortcomings of evaluation metrics
1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC
YAGO3-TC Dataset

What is Our goal?

• Create a benchmark that align with real-world application
• Properly differentiate between models
• Capture reasoning powers

What are existing challenges?

• The knowledge graphs are not complete
• There so many non existent links
• Identifying the factuality of missing information is hard
YAGO3-TC Creation

1. Randomly choose a subset of YAGO3-10 test

2. Identify top scoring triples from the models that are unknown to be true

3. Filter triples

4. Crowdsourcing pipeline
Crowdsourcing Pipeline

Initial Annotations (Round 1)

Carles Puyol plays for?
- Catalonia National Team  ✓ ✓ ✗
- Barcelona  ✓ ✓ ✓
- Real Madrid  ✓ ✗ ✗
- Man United  ✗ ✗ ✗

Selective Reannotation (Round 2)

✓ Catalonia National Team  ✓ Barcelona  ✗ Man United  ✗ Real Madrid

- ~ 30 K triples
- ~ 10% positives
Evaluation Using YAGO3-TC

- Triple classification

  ![Graph showing ACC and F1 for different models]

  - SOTA models perform poorly
  - Huge drop in accuracy

- Calibration

  ![Graph showing ratio of positives vs. mean score]

  - Reverse order of models
  - Overconfident
Discussion

• Ranking metrics are not very trustworthy
• Triple classification is not robust
• Real-world adoption of KG needs better evaluation techniques
• YAGO3-TC is the first step toward this goal

We propose a web-hosted evaluation platform to update YAGO3-TC using new KG completion models
Revisiting Evaluation of KB Completion

Evaluating KG Completion

Shortcomings

1. Semi-Inverse relations
2. Calibration
3. Triple classification robustness

Introducing YAGO3-TC

Thank You!

Website (code, data and leaderboard):

pouyapez.github.io/yago3-tc/

Contact me: pezeshkp@uci.edu